COUPLING AND MONOTONICITY OF QUEUEING PROCESSES

EVSEY MOROZOV

Abstract. The main purpose of this work is to give a survey of main monotonicity properties of queueing processes based on the coupling method. The literature on this topic is quite extensive, and we do not consider all aspects of this topic. Our more concrete goal is to select the most interesting basic monotonicity results and give simple and elegant proofs. Also we give a few new (or revised) proofs of a few important monotonicity properties for the queue-size and workload processes both in single-server and multi- server systems. The paper is organized as follows. In Section 1, the basic notions and results on coupling method are given. Section 2 contains known coupling results for renewal processes with focus on construction of synchronized renewal instants for a superposition of independent renewal processes. In Section 3, we present basic monotonicity results for the queue-size and workload processes. We consider both discrete- and continuous-time queueing systems with single and multi servers. Less known results on monotonicity of queueing processes with dependent service times and interarrival times are also presented. Section 4 is devoted to monotonicity of general Jackson-type queueing networks with Markovian routing. This section is based on the notable paper [17]. Finally, Section 5 contains elements of stability analysis of regenerative queues and networks, where coupling and monotonicity results play a crucial role to establish minimal sufficient stability conditions. Besides, we present some new monotonicity results for tandem networks.

1. Introduction

Coupling is a common way to present random variables, generally defined on different probability spaces, as elements of a new common probability space keeping their predetermined (marginal) distributions. Another important aspect we discuss in the work is a coupling time of two renewal processes defined on a common probability space and governed by the same

The work is supported by Russian Foundation for Basic Research under grant 07-07-00088.

interrenewal distribution. We show how to couple different versions of renewal processes in particular, delayed renewal process with its stationary version or with its zero-delayed version.

The concept of coupling is also very useful both in estimation the (steady-state) performance measure and the obtaining the rate of convergence to stationarity. The most interesting aspect of coupling we consider in this work is its application to stability analysis obtained via monotonicity properties.

The most important sources on coupling method are [1, 6, 22], and other sources for the survey are [17, 18, 25].

Consider two stochastic processes $X = \{X_t, t \geq 0\}$, $X' = \{X'_t, t \geq 0\}$ with distributions P, P' respectively, with the same state space (E, B) (and defined in general on different probability spaces) with the same parameter $t \in N = \{0, 1, \ldots\}$ or $t \in \mathsf{R}_+ = [0, \infty)$. *The coupling* of X and X' is a realization $\hat{X} = (\hat{X}, \hat{X}')$ on a common probability space with state space $(E^2, \mathsf{B} \otimes \mathsf{B})$ such that

$$\hat{X} =_{st} X, \ \hat{X}' =_{st} X' \tag{1.1}$$

($=_{st}$ means equality in distribution). In other words, distributions of random elements X, X' are the marginals of the distribution of coupling $\hat{X}$. Coupling allows a sample-path comparison of realizations of random processes. In what follows we (as a rule) assume that the process $\tilde{X} = (X, X')$ is already a coupling with a distribution P.

Assume original processes have the same distributions, or the process X could be obtained by a "shift" of the process X'. Then the following notion is used.

Definition. Let $\tilde{X} = (X, X')$ be a coupling. A random time $T \in [0, \infty]$ is a *coupling time* of $\tilde{X}$ if

$$X_t = X'_t \quad \text{for} \quad t \geq T \text{ on } \{T < \infty\}. \tag{1.2}$$

The coupling is called *successful* if $\mathsf{P}(T < \infty) = 1$. Note that for any measurable set A,

$$\begin{aligned} &|\mathsf{P}(X_t \in A) - \mathsf{P}(X'_t \in A)| = \\ &= \ |\mathsf{P}(X_t \in A, T \leq t) - \mathsf{P}(X'_t \in A, T \leq t| \\ &+ \ |\mathsf{P}(X_t \in A, T > t) - \mathsf{P}(X'_t \in A, T > t| \leq \mathsf{P}(T > t), \end{aligned}$$

where we use (1.2). Now taking supremum over A, we obtain the main *coupling inequality*

$$||\mathsf{P}(X_t \in \cdot) - \mathsf{P}(X'_t \in \cdot)|| \leq \mathsf{P}(T > t), \tag{1.3}$$

estimating the total variation distance between distributions X and X'. In fact, this inequality holds in sample-path sense if we replace values X_t, X'_t

by shifted versions $\theta X_t =: (X_{t+s},\, s \geq 0)$ and $\theta X'_t =: (X'_{t+s},\, s \geq 0)$, respectively, because since instant T both processes coincide. Inequality (1.3) allows to obtain convergence rate to stationarity if say X' is a strictly stationary version of X with $\mathsf{P}'(X'_t \in \cdot) = \pi(\cdot)$ for all t. If there is a function $\phi(x) \to \infty$ such that $\mathsf{E}\phi(T) < \infty$, then the following rate of convergence in total variation holds:

$$||\mathsf{P}(X_t \in \cdot) - \pi(\cdot)|| \leq \mathsf{P}(T > t) \leq \frac{\mathsf{E}\phi(T)}{\phi(T)} = O(\frac{1}{\phi(t)}).$$

In the context of this work, it is especially important to discuss a special case of the coupling leading to stochastic ordering. Let X, Y be real-valued random variables with distributions F_X, F_Y. Then $X \leq_{st} Y$ (in the sense of *stochastic ordering*) if $F_X(x) \geq F_Y(x)$ for all x. (It is denoted $F_X \geq F_Y$.)

This definition is equivalent to the following: i) $\mathsf{E}f(X) \leq \mathsf{E}f(Y)$ for all increasing functions $f : \mathsf{R} = (-\infty, \infty) \to \mathsf{R}$ or ii) there exist random variables X', Y' (defined on a common probability space) such that $X' =_{st} X$, $Y' =_{st} Y$ and $X' \leq Y'$ with probability 1 (w.p.1).

Condition ii) allows us to compare stochastic processes in a sample -path sense given ordering between distributions. To construct a common probability space (supporting new random variables with the same marginal distributions), the quantile functions are used. We describe this construction for the state space $E = \mathsf{R}$.

Consider random variables X, Y with distributions F_X, F_Y. Define quantile functions Q_X, Q_Y as

$$Q_X(u) = \inf(x :\, F_X(x) \geq u),\; Q_Y(u) = \inf(x :\, F_Y(x) \geq u),\; u \in (0,1). \tag{1.4}$$

It is easy to see that

$$Q_X(u) \leq x \text{ iff } F_X(x) \geq u, \quad Q_Y(u) \leq x, \text{ iff } F_Y(x) \geq u. \tag{1.5}$$

To construct a common probability space, we take $\Omega^* = (0,\, 1)$ with Borel sigma-algebra $\mathtt{B}^*$ and Lebesgue measure $\mathsf{P}^*(dx) = dx$. In other words, elements $\omega^* \in \Omega^*$ are uniformly distributed in $(0, 1)$. Define new variables X^*, Y^* as

$$X^*(\omega^*) = Q_X(\omega^*),\; Y^*(\omega^*) = Q_Y(\omega^*),\; \omega^* \in \Omega^*. \tag{1.6}$$

Then by (1.5),

$$\begin{aligned}
\mathsf{P}^*(\omega^* : X^*(\omega^*) \leq x) &= \mathsf{P}^*(\omega^* : Q_X(\omega^*) \leq x) \\
&= \mathsf{P}^*(\omega^* : \omega^* \leq F_X(x)) = F_X(x); \\
\mathsf{P}^*(\omega^* : Y^*(\omega^*) \leq x) &= \mathsf{P}^*(\omega^* : Q_Y(\omega^*) \leq x) \\
&= \mathsf{P}^*(\omega^* : \omega^* \leq F_Y(x)) = F_Y(x).
\end{aligned}$$

Thus, random variables X^*, Y^* defined on the common probability space $(\Omega^*, \mathsf{B}^*, \mathsf{P}^*)$ have given (marginal) distributions F_X, F_Y, respectively. Hence, an order $F_X \geq (\leq) F_Y$ between given distributions implies P^*-a.s. inequalities between new random variables

$$X^*(\omega^*) = Q_X(\omega^*) \leq (\geq) Y^*(\omega^*) = Q_Y(\omega^*), \quad \omega^* \in \Omega^*. \tag{1.7}$$

2. Coupling of renewal processes

Following [1], we construct a successful coupling of a renewal process with its stationary version. This construction is used in stability analysis in Section 5.

Consider a renewal (i.i.d.) sequence X_n, $n \geq 1$, with $X_n > 0$, with mean $\mathsf{E}X = \mu < \infty$ (where X is a generic renewal period) and with *spread-out distribution* F. It means that F^{*n}, n-convolution F with itself, has an absolutely continuous component for some $n \geq 1$. Define $S_0 = 0$ and random walk $S_n = X_1 + \cdots + X_n$, $n \geq 1$. Define also the number of renewals in $(0, t]$ as

$$N(t) = \#\{n \geq 1 : S_n \leq t\}, \ t \geq 0. \tag{2.1}$$

By the Stone's decomposition [19], the renewal function $U(t) = \mathsf{E}N(t)$ can be written as

$$U = U_1 + U_2, \tag{2.2}$$

where U_i are nonnegative measures on $[0, \infty)$, measure U_2 is bounded (total variation $||U_2|| < \infty$) and there exists a bounded continuous density for measure U_1,

$$u_1(x) = dU_1(x)/dx \to \frac{1}{\mu} \text{ as } x \to \infty. \tag{2.3}$$

For each $t \geq 0$ we define the *forward renewal time (overshoot)*

$$B_t = \inf_{n \geq 0} (S_n - t : S_n - t \geq 0).$$

In particular, $B_0 = 0$. Define distribution $G_t(x) = \mathsf{P}(B_t \leq x)$ and the density of the stationary forward renewal time,

$$F_0(dx) = \frac{1}{\mu}(1 - F(x))dx. \tag{2.4}$$

The following statements are adopted from [1].

Theorem 2.1.

$$||G_t - F_0|| \to 0 \quad \text{for any} \quad X_1 \tag{2.5}$$

if and only if F is spread-out.

We recall the renewal equation

$$Z(t) = z(t) + \int_0^t Z(t-u)dF(u),\ t \geq 0, \tag{2.6}$$

where Z is unknown function on $[0, \infty)$ and z is a known one. In what follows we will assume (without loss of generality) that

$$F(dx) \geq cdx \ \text{ if } \ x \in (a,\ a+2b) \tag{2.7}$$

for some $a,\ b > 0,\ c > 0$.

Lemma 2.1. *If z is bounded on finite intervals then solution to renewal equation (2.6) is*

$$Z(t) = \int_0^t z(t-u)dU(u). \tag{2.8}$$

We include an instructive proof of the following statement.

Theorem 2.2. *For the zero-delayed case (that is X_1 has the same distribution as others), distribution G_t has a common uniform component on $(0, b)$ for all $t \geq C$ where C is a finite constant. In other words, for some $\delta \in (0, 1)$,*

$$\mathsf{P}(u < B_t \leq v) \geq \delta \frac{v-u}{b},\ \ 0 < u < v < b,\ t \geq C. \tag{2.9}$$

Proof [1]. Denote $X = X_1$ and fix arbitrary $u < v$. Let $Z(t) = \mathsf{P}(u < B_t \leq v)$ and

$$z(t) = \mathsf{P}(u < B_t \leq v,\ X > t) = \mathsf{P}(B_t \leq v,\ X > t) - \mathsf{P}(B_t \leq u,\ X > t).$$

Because $\mathsf{P}(B_t \leq v,\ X > t) = F(t+v)$ then it follows that

$$z(t) = F(t+v) - F(t+u). \tag{2.10}$$

By Stone's decomposition, $U \geq U_1$, and then solution $Z(t)$ to equation (2.6) gives for $t \geq a + b$ (see (2.3), (2.8))

$$Z(t) \geq \int_0^t z(y)u_1(t-y)dy \geq \int_a^{a+b} z(y)u_1(t-y)dy. \tag{2.11}$$

If $y \in (a,\ a+b)$ and $0 < u < v < b$ then $a < y + u < y + v < a + 2b$, and we obtain from (2.6), (2.10) that

$$z(y) = \int_{x=y+u}^{y+v} dF(x) \geq c(v-u). \tag{2.12}$$

By (2.3), there exists t_0 such that

$$u_1(t) \geq \frac{1}{2\mu}, \quad t \geq t_0.$$

Then it follows from (2.11), (2.12) that for $t \geq t_0 + a + b$,

$$\int_a^{a+b} z(y)u_1(t-y)dy \geq c(v-u)\frac{b}{2\mu} = \frac{v-u}{b} \cdot \frac{cb^2}{2\mu}. \tag{2.13}$$

By (2.7),

$$\mu = \mathsf{E}X \geq \int_a^{a+2b} xdF(x) \geq c\frac{(a+2b)^2 - a^2}{2} = 2bc(a+b),$$

and we obtain $2\mu \geq 4b^2c$. Hence, (2.9) holds with $\delta = \frac{cb^2}{2\mu} \leq 1/4$ and with any $C \geq t_0 + a + b$. ■

Above given construction is well-known and allows us to construct a (less known) coupling of a few renewal processes. Consider two renewal processes, zero-delayed and stationary, with the forward renewal times B_t, B'_t, respectively. Let C be fixed and satisfy (2.9). Define

$$t_0 = 0,\ B_{t_0} = 0,\ L_0 = \max(B_{t_0},\ B'_{t_0}),\ t_1 = t_0 + L_0 + C,$$

where B'_{t_0} has distribution (2.4). We choose random variables U_k, V_k such that

$$\mathsf{P}(U_k = 1) = 1 - \mathsf{P}(U_k = 0) = \delta \in (0,1),$$

and that V_k are uniform on $(0, b)$. Also let

$$M_0 = L_0 + C - B_{t_0},\ \ M'_0 = L_0 + C - B'_{t_0},$$

and define for any $k \geq 1$,

$$L_k = \max(B_{t_k},\ B'_{t_k}),\ t_{k+1} = t_k + L_k + C,\ M_k = L_k + C - B_{t_k},$$
$$M'_k = L_k + C - B'_{t_k}.$$

First of all we have

$$M_k \geq C,\ \ M'_k \geq C,\ \ t_{k+1} - t_k = C + L_k \geq \max(M_k,\ M'_k).$$

A key observation is that the instants $t_{k+1} - M_k = t_k + B_{t_k}$ and $t_{k+1} - M'_k = t_k + B'_{t_k}$ are the renewal points of the corresponding processes. Now we put

$$B_{t_{k+1}} = U_kV_k + (1-U_k)R_k,\ \ B'_{t_{k+1}} = U_kV_k + (1-U_k)R'_k,$$

where variables R_k, R'_k are chosen with the *rest distributions*

$$\frac{\mathsf{P}(B_{t_k} \leq x) - \min(x,b)\delta/b}{1-\delta},\ \ \frac{\mathsf{P}(B'_{t_k} \leq x) - \min(x,b)\delta/b}{1-\delta},$$

respectively. Then it follows from Theorem 2.2 that $B_{t_{k+1}}$, $B'_{t_{k+1}}$ have the same distributions as M_k, M'_k, respectively. At that, variables U_k, V_k, R_k, R'_k are taken independent of all preceding U_r, V_r, R_r, R'_r.

The renewals for the zero-delayed process $\{S_n\}$ in interval $[t_k - M_k, t_k]$ are constructed using the conditional distribution given that its forward renewal time at instant M_k has the constructed value $B_{t_{k+1}}$ (similarly, for the stationary process $\{S'_n\}$). The procedure is stopped at the instant $\sigma = \inf(n \geq 0 : U_n = 1)$, when both processes have a common renewal at instant $T = t_\sigma + L_{\sigma+1}$. Indeed, if $U_n = 1$ then overshoots $B_{t_{k+1}}$, $B'_{t_{k+1}}$ are uniformly distributed in $(0, b)$ and, by coupling, we obtain the common renewal point $t_{k+1} + B_{t_{k+1}} = t_{k+1} + B'_{t_{k+1}} = t_{k+1} + V_k$. Then we construct a new renewal process $\{\tilde{S}_n\}$ (coupling of two original processes) which has the same renewals as $\{S_n\}$ before T and the same renewals as $\{S'_n\}$ after T, and this new process has required marginal distribution. The stopping time σ has geometrical distribution $\mathsf{P}(\sigma = n) = \delta(1-\delta)^n$, so $\sigma < \infty$ and $T < \infty$ (w.p.1).

A slight modification of this construction allows us to obtain common renewal points for a superposition of m independent renewal processes, see [18]. Let $m = 2$ and F_i be the interevent distribution of the process $i = 1, 2$. Assume $F_1(dx) \geq c dx,\ x \in (a, a + 2b)$, for suitable constants $a, b > 0, c > 0$. Let $B_t^{(i)}$ be the forward renewal time of the process i at instant t; ν be the uniform measure on $(0, b)$ and let $t_n^{(i)}$, $n \geq 1$, be the renewal instants of the process $i = 1, 2$. Consider the i.i.d. sequence V_i, $i \geq 1$, with distribution ν and independent of the i.i.d. $0-1$ random variables U_i, $n \geq 1$, with $\mathsf{P}(U_1 = 1) = \delta$. Introduce the overshoot process $B_t = (B_t^{(1)}, B_t^{(2)})$, $t \geq 0$. Then by Theorem 2.2,

$$\mathsf{P}(B_t^{(1)} \in A) \geq \delta \nu(A)$$

for any set $A \in \mathtt{B}(0, b)$ and $t \geq C$. Let

$$t_0 = 0,\ L_0 = B_{t_0}^{(1)},\ t_1 = \min(t_n^{(2)} : t_n^{(2)} \geq t_0 + L_0 + C).$$

Then we take $B_{t_1}^{(1)} = U_1 V_1 + (1 - U_1) R_1$ and define recursively

$$\begin{aligned} t_k &= \min(t_n^{(2)} : t_n^{(2)} \geq t_{k-1} + L_{k-1} + C), \\ B_{t_k}^{(1)} &= U_k V_k + (1 - U_k) R_k,\ L_k = B_{t_k}^{(1)},\ k \geq 1, \end{aligned}$$

with U_k, V_k, R_k independent of all preceding U_l, V_l, R_l. Let $T_0 = 0$ and $T_{n+1} = \min(t_k > T_n : U_k = 1)$, $n \geq 0$. It then follows that $\{T_n\} \in \{t_n^{(2)}\}$, and the overshoot $B_{T_n} = (B_{T_n}^{(1)}, B_{T_n}^{(2)})$ at each instant T_n has distribution $\nu \otimes F_2$ and is independent of the pre-history B_t, $t < T_{n-1}$. As a result,

$\{T_n\}$ are regeneration points for the *one-dependent process* B_t, $t \geq 0$. (This process has the i.i.d periods $\{T_{n+1} - T_n\}$ and *one-dependent cycles*, see [18].)

To couple $m \geq 2$ renewal processes, we assume that at least $m-1$ interevent distributions $F_2, \ldots, F_m$ say are spread-out. Then we replace L_n by $\max_{2 \leq i \leq m} B_{t_n}^{(i)}$, constant C by $\max_{2 \leq i \leq m} C_i$ and the event $\{U_n = 1\}$ by the event $\{\min_{2 \leq i \leq m} U_n^{(i)} = 1\}$, where constant C_i, overshot $B_{t_n}^{(i)}$ and variable $U_n^{(i)}$ relate to the ith renewal process. (For more details see [18]).

We mention *splitting* of a Markov chain $\{X_n,\, n \geq 0\}$ with the state space (R, B) and transition kernel $\mathsf{P}^n(x, \cdot)$, $n \geq 1$ [1]. Consider a set A and let $\tau(A) = \inf(n \geq 1 : X_n \in A)$. The set A is *recurrent* if $\mathsf{P}_x(\tau(A) < \infty) = 1$ (for all initial states x) and is *regenerative* if in addition,

$$\mathsf{P}^r(x, \cdot) \geq \varepsilon \lambda(\cdot),\ x \in A, \tag{2.19}$$

for some $r \geq 1$, $\varepsilon \in (0, 1)$ and a probability measure λ. If a regenerative set exists, then the chain is called *Harris recurrent*, and one can construct an embedded renewal process of regenerations as follows. We realize usual version of the process up to the time $\tau(A)$. Then, with probability ε, we realize $X_{\tau(A)+r}$ according to distribution λ, and thus a regeneration occurs at instant $\tau(A) + r$. Otherwise, with probability $1 - \varepsilon$, we use the rest distribution,

$$\frac{\mathsf{P}^r(X_{\tau(A)}, \cdot) - \varepsilon \lambda(\cdot)}{1 - \varepsilon}. \tag{2.20}$$

Then we realize fragment $\{X_{\tau(A)+k},\, 0 < k < r\}$ in according to the conditional distribution of the process $\{X_k,\, 0 < k < r\}$ given boundary values $X_0 = X_{\tau(A)}$, $X_r = X_{\tau(A)+r}$. The procedure is repeated with the new initial value $X_{\tau(A)+r} = x$, and so on. As a result we get a new Markov chain with the same transition probabilities which is *one-dependent regenerative with dependent adjacent cycles and independent regeneration periods*. As above, the coupling time has geometrical distribution.

3. Coupling and monotonicity of queues

3.1. **Classical systems.** In this section, we collect together the most important and interesting applications of the coupling in analysis of queues. The main purpose is to find monotonicity properties of queues and use them for comparison of queueing performance measures. The basic sources are [1, 2, 4, 25].

First of all we will follow [4] to obtain monotonicity property for queue-size and workload both in continuous and in discrete time. Our proofs are mainly modified and simplified versions of that are given in [4].

We consider two $GI/G/m$ queues Q, $\tilde{Q}$ with $m \geq 1$ servers. Let A, B be interarrival time and service time distributions in queue Q, respectively. (Corresponding variables in queue $\tilde{Q}$ we endow with tildes.) Let t_n be arrival instants and $\tau_n = t_{n+1} - t_n$, $n \geq 1$, be the i.i.d.interarrival times. (In the delayed case, τ_1 has in general another distribution.) Denote the i.i.d. service times S_n, $n \geq 1$.

To simplify notations, we will not often distinguish original and coupled variables living in a common space. For instance, if distributions $F_X \geq F_Y$ then we write $X \leq Y$ instead of $X^* \leq Y^*$ (P^*-a.s.) for the coupled variables X^*, Y^* such that $X^* =_{st}$, $Y^* =_{st} Y$ (see (1.7)). Also let S and τ denote generic service time and input interval, respectively. Let $W_i(t)$ be the (left-continuous) ith smallest unfinished workload among all m servers at instant t, and $W(t) = (W_1(t), \ldots, W_m(t))$. Introduce backward interarrival time at instant t, $\tau(t) = \max_n(t - t_n : t - t_n > 0)$. Note that the process $\tau(t)$, $t \geq 0$, is right-continuous and $\tau(t_{n+1}) = \tau_n$. Let also $N(t) = \max(n \geq 1 : t_n < t)$ be the number of arrivals in $(0, t)$. Unlike definition of the process N in section 2, here $N(\cdot)$ is left-continuous, $N(0) = 0$ and $N(t_n) = n - 1$. Denote $W_i(t_n) = W_n^{(i)}$, $i = 1, \ldots, m$. We note the following relation connecting processes $W(t)$ and W_n:

$$W(t) = R(W_{N(t)}^{(1)} + S_{N(t)} - \tau(t),\, W_{N(t)}^{(2)} - \tau(t), \ldots, W_{N(t)}^{(m)} - \tau(t))^+, \quad (3.1)$$

where operator $(\cdot)^+$ is component wise and operator R puts components in an increasing order. For $t = t_{n+1}$ relation (3.1) transforms into well-known Kiefer-Wolfowitz recursion for the workload vector $W_n = (W_n^{(1)}, \ldots, W_n^{(m)})$:

$$\begin{aligned} W_{n+1} &= R(W_n^{(1)} + S_n - \tau_n,\, W_n^{(2)} - \tau_n, \ldots, W_n^{(m)} - \tau_n)^+ \\ &= R(W_n + eS_n - \mathbf{I}\tau_n)^+,\ n \geq 1, \end{aligned} \quad (3.2)$$

where vectors $e = (1, 0, \ldots, 0)$, $\mathbf{I} = (1, \ldots, 1) \in R_+^n$. Now we prove monotonicity of the workload process $W = (W_n,\, n \geq 1)$. Assume that $W_n \leq \tilde{W}_n$ and that

$$\tilde{S}_n \leq S_n,\ \tau_n \leq \tilde{\tau}_n,$$

for some n. Then obviously,

$$(\tilde{W}_n + e\tilde{S}_n - \mathbf{I}\tilde{\tau}_n)^+ \leq (W_n + eS_n - \mathbf{I}\tau_n)^+,\ n \geq 1.$$

Assume that m-dimensional vectors X, Y are ordered, $X \leq Y$, and show that operator R keeps the ordering, that is

$$RX =: X^* = (X_1^*, \ldots, X_m^*) \leq RY =: Y^*. \quad (3.3)$$

This implies the monotonicity of mapping (3.2) with respect to (w.r.t) input interval (decreasing) and service time (increasing). Define index $n(i)$ as

$$X_i^* = X_{n(i)},\ i = 1, \ldots, m.$$

Since $X_{n(1)} = \min_k(X_k)$ then, by (3.3), the inequality

$$X_1^* =: X_{n(1)} > Y_1^* =: Y_{n(1)}$$

is impossible and thus $X_1^* \le Y_1^*$. Also inequality $X_2^* =: X_{n(2)} > Y_{n(2)} = Y_2^*$ is impossible since $X_{n(2)} = \Theta_2(X_k,\, k = 1, \ldots, m)$, where operator Θ_n selects the nth smallest element. These arguments show that operator R keeps (component-wise) order. It immediately gives us the following result.

Theorem 3.1. *Assume that* $\tilde{W}_1 \le W_1$ *w.p.1 and*

$$\tilde{\tau} \ge_{st} \tau,\ \tilde{S} \le_{st} S. \tag{3.4}$$

Then the following (component-wise) ordering holds:

$$\tilde{W}_n \le_{st} W_n,\ n \ge 1. \tag{3.5}$$

In particular, waiting times are ordered as

$$\tilde{W}_n^{(1)} \le_{st} W_n^{(1)},\ n \ge 1. \tag{3.6}$$

In what follows, we write down stochastic relations for original variables as $X = (\le, \ge)Y$ (instead of $=_{st}$ $(\le_{st}, \ge_{st})$) although such (w.p.1) relations indeed hold for the coupled variables.

The following statement is a revised and extended version of Theorem 3.1 from [4]. Let, in the system Q, ν_n be the number of customers just before arrival instant t_n; D_n be the departure instant of customer n; B_n be the nth service beginning epoch, and Q_n be the number of customers waiting in the queue just before t_n.

Theorem 3.2. *If* $\nu_1 = \tilde{\nu}_1 = 0$, $\tau = \tilde{\tau}$, $S \ge \tilde{S}$, *then*

$$\tilde{\nu}_n \le \nu_n,\ \tilde{Q}_n \le Q_n,\ n \ge 1. \tag{3.7}$$

Proof. Using coupling we have $t_n = \tilde{t}_n$, $n \ge 1$. Then by (3.6),

$$D_n = t_n + W_n^{(1)} + S_n \ge t_n + \tilde{W}_n^{(1)} + \tilde{S}_n = \tilde{D}_n,\ n \ge 1. \tag{3.8}$$

Introduce the number of departures in interval $[0,\, t_n]$ in both systems:

$$\Delta(n) = \#\{k : D_k \le t_n\},\ \tilde{\Delta}(n) = \#\{k : \tilde{D}_k \le \tilde{t}_n\}. \tag{3.9}$$

It follows from (3.8) that $\tilde{\Delta}(n) \ge \Delta(n)$ and we obtain the 1st inequality (3.8):

$$\nu_n = n - 1 - N(n) \ge n - 1 - \tilde{N}(n) = \tilde{\nu}_n,\ n \ge 1. \tag{3.10}$$

Because

$$B_n = \max(t_n,\, D_{n-m}),\ n \ge 1\ (D_k = 0 \text{ for } k < 0), \tag{3.11}$$

then by (3.8),

$$B_n \ge \tilde{B}_n, \tag{3.12}$$

and thus,

$$M(n) = \#\{k : B_k < t_n\} \le \tilde{M}(n) = \#\{k : \tilde{B}_k < t_n\}.$$

It implies desired inequality for the queue size,

$$Q_n = n - 1 - M(n) \ge n - 1 - \tilde{M}(n) = \tilde{Q}_n,\ n \ge 1.$$

■

Note 3.1. It is possible to extend Theorem 3.2 for non-empty initial conditions, when $\tilde{\nu}_1 \le \nu_1$ and initial unfinished service times are ordered in an evident way.

Let ξ_∞, resp., $\xi(\infty)$ denote a weak limit of the sequence ξ_n, resp., $\xi(t)$. Obviously, inequalities (3.5)- (3.7) hold also for weak limits (if exist) that is (stochastically)

$$\tilde{W}_\infty \le W_\infty,\ \tilde{W}^{(1)}_\infty \le W^{(1)}_\infty,\ \tilde{\nu}_\infty \le \nu_\infty,\ \tilde{Q}_\infty \le Q_\infty. \tag{3.13}$$

Let $\nu(t)$ be the number of customers in the system Q at instant t^+. Now we present a simpler proof of the following statement which includes theorems 4.1 and 4.2 from [4].

Theorem 3.3. *If $\nu(0) = \tilde{\nu}(0) = 0,\ \tau = \tilde{\tau}$ and $S \ge \tilde{S}$ then*

$$\nu(t) \ge \tilde{\nu}(t),\ W(t) \ge \tilde{W}(t),\ t \ge 0. \tag{3.14}$$

In particular, virtual waiting times are ordered as

$$W_1(t) \ge \tilde{W}_1(t),\ t \ge 0.$$

Proof. To establish first inequality in (3.14), we note that by (3.8),

$$\tilde{\Lambda}(t) =: \#\{k : \tilde{D}_k \le t\} \ge \#\{k : D_k \le t\} =: \Lambda(t),\ t \ge 0.$$

Then the desired result follows since

$$\nu(t) = N(t) - \Lambda(t) \ge \tilde{N}(t) - \tilde{\Lambda}(t) = \tilde{\nu}(t),$$

where $N(t) = \tilde{N}(t)$ is the number of arrivals in $(0, t]$. (Note that $\Delta(n) = \Lambda(t_n)$ and $\tilde{\Delta}(n) = \tilde{\Lambda}(t_n)$, see (3.9).) The 2nd inequality in (3.14) is a direct consequence of the monotonicity of Kiefer-Wolfowitz recursion (3.2) and its continuous-time analog (3.1), because $S_{N(t)} =_{st} S$, $\tau(t) =_{st} \tilde{\tau}(t)$ and $\tilde{W}_{\tilde{N}(t)} \le_{st} W_{N(t)}$ by (3.5). ■

In the next two statements, we simplify and modify the proof of three results from [4] for a single-server system $GI/G/1$. We assume that conti-

nuous- and discrete- time processes in systems Q, $\tilde{Q}$ have (in evident notations) weak limits, $W(\infty)$, $\nu(\infty)$, W_∞, ν_∞ and $\tilde{W}(\infty), \tilde{\nu}(\infty)$, $\tilde{W}_\infty$, $\tilde{\nu}_\infty$, respectively. Sufficient conditions for the discrete-time limits to exist are:

$$\rho =: \mathsf{E}\tau/\mathsf{E}S < 1,\ \tilde{\rho} =: \mathsf{E}\tilde{\tau}/\mathsf{E}\tilde{S} < 1. \tag{3.15}$$

These conditions in turn imply conditions $\mathsf{P}(\tau > S) > 0$ and $\mathsf{P}(\tilde{\tau} > \tilde{S}) > 0$, respectively. These conditions also imply aperiodicity of the (discrete-time) regeneration period and finiteness of its mean. If moreover, distributions of τ, $\tilde{\tau}$ are non-lattice then regeneration period is also non-lattice and weak limits for continuous time also exist [1].

Theorem 3.4. *Assume that*

$$\tau \le \tilde{\tau},\ S = \tilde{S}. \tag{3.16}$$

Then regardless of the initial state,

$$\tilde{W}(\infty) \le W(\infty),\ \tilde{\nu}(\infty) \le \nu(\infty). \tag{3.17}$$

Proof. By (3.15),(3.16), $1 > \rho \ge \tilde{\rho}$ and, by (3.6), also $W_n = W_n^{(1)} \ge \tilde{W}_n^{(1)} = \tilde{W}_n$ and hence, $W_\infty \ge \tilde{W}_\infty$. As in [4] we now use explicit relations between limit distributions. Denote $\hat{S}$ the stationary service time with (common for both systems) distribution

$$\mathsf{P}(\hat{S} \le x) = \frac{1}{\mathsf{E}S}\int_0^x (1 - B(y))dy,$$

where B is the distribution of service time. Then by (3.16) and [21],

$$\begin{aligned} \mathsf{P}(\nu(\infty) > k) &= \rho\mathsf{P}(\tau_1 + \cdots + \tau_k \le W_\infty + \hat{S}) \\ &\ge \tilde{\rho}\mathsf{P}(\tilde{\tau}_1 + \cdots + \tilde{\tau}_k \le \tilde{W}_\infty + \hat{S}) \\ &= \mathsf{P}(\tilde{\nu}(\infty) > k), \end{aligned}$$

and the 2nd relation in (3.17) follows. Also the 1st inequality in (3.17) follows since for each $x > 0$,

$$\mathsf{P}(W(\infty) > x) = \rho\mathsf{P}(W_\infty + \hat{S} > x) \ge \tilde{\rho}\mathsf{P}(\tilde{W}_\infty + \hat{S} > x) = \mathsf{P}(\tilde{W}(\infty) > x),$$

where the equalities have been proved in [21], Theorem 2. ■

Let $Q(t)$ be the number of customers waiting in the queue Q at instant t and $Q(\infty)$ be its weak limit (if exist).

Theorem 3.5. *Let, in a stationary queue $GI/M/k$, assumptions (3.16) hold. Then*

$$\tilde{Q}(\infty) \le Q(\infty).$$

Proof. It is proved in [20] that $P(Q(\infty) \geq n) = \frac{\rho}{k}\mathsf{P}(\nu_\infty \geq n+k)$ for any $n \geq 0$. Because $\rho \geq \tilde{\rho}$ by (3.16), and $\nu_\infty \geq \tilde{\nu}_\infty$ by Theorem 3.2, we obtain

$$\frac{\rho}{k}\mathsf{P}(\nu_\infty \geq n+k) \geq \frac{\tilde{\rho}}{k}\mathsf{P}(\tilde{\nu}_\infty \geq n+k) = \mathsf{P}(\tilde{Q}(\infty) \geq n),\ n \geq 0.$$

■

The proof of the following result in [4] is not satisfactory, and we give another proof.

Theorem 3.6. *Let, in m-server queues Q, $\tilde{Q}$,*

$$\tilde{W}_1 \leq W_1,\ \tau \leq \tilde{\tau},\ \tilde{S} = S. \tag{3.18}$$

Then

$$\tilde{\nu}_n \leq \nu_n,\ n \geq 1. \tag{3.19}$$

Proof. By (3.18) and Theorem 3.1, $\tilde{W}_n \leq W_n$, $n \geq 1$. It also follows from (3.18) that the difference $\Delta_n =: \tilde{t}_n - t_n \geq 0$ is non-decreasing in n. Hence, the departure instants of customer n in both systems satisfy inequality

$$\tilde{T}_n =: \tilde{t}_n + \tilde{W}_n^{(1)} + S_n \leq t_n + \Delta_n + W_n^{(1)} + S_n =: \Delta_n + T_n,\ n \geq 1.$$

Because

$$\tilde{T}_k \leq T_k + \Delta_k \leq T_k + \Delta_n,\ n \geq k \geq 1,$$

then we obtain

$$\begin{aligned} C(n) &= \#\{k : T_k < t_n\} = \#\{k : T_k + \Delta_n < t_n + \Delta_n\} \\ &\leq \#\{k : T_k + \Delta_k < \tilde{t}_n\} \leq \#\{k : \tilde{T}_k < \tilde{t}_n\} = \tilde{C}(n). \end{aligned} \tag{3.20}$$

Hence,

$$\nu_n = n - 1 - C(n) \geq n - 1 - \tilde{C}(n) = \tilde{\nu}_n, \tag{3.21}$$

and (3.19) follows. ■

Note that (3.18) does not imply ordering (3.19) for continuous-time queue-size process.

Now we compare continuous-time workload processes in an m-server $GI/G/m$ queue Q with the i.i.d. service times S_n, $n \geq 1$, with a single-server queue Q^* in which service times S_n are reduced in m times. We will keep previous notations assuming $t_1 = 0$. Let $W_t = \sum_{i=1}^m W_i(t)$ be the total workload at instant t^-. As always, we consider a coupling of queues Q, Q^* on the same probability space, with the same interarrival times $\tau_n^* = \tau_n$ and service times $S_n^* = S_n/m$, $n \geq 1$. Let W_t^* be the workload at instant t in queue Q^*.

Theorem 3.7 [1]. *If* $mW_0^* \le W_0$ *then*

$$mW_t^* \le W_t. \tag{3.22}$$

Proof. Using inequality $(x+y)^+ \le x^+ + y^+$ m times we have for $0 = t_1 \le t < \tau_1 = t_2$:

$$\begin{aligned} mW_t^* &= m(W_0^* + S_1/m - t)^+ = (mW_0^* + S_1 - tm)^+ \\ &\le (\sum_{i=1}^m W_i(0) + S_1 - tm)^+ = (W_1(0) + S_1 - t + \sum_{i=2}^m (W_i(0) - t))^+ \\ &\le (W_1(0) + S_1 - t)^+ + \sum_{i=2}^m (W_i(0) - t)^+ = W_t. \end{aligned}$$

Since $mW_{t_2}^* \le W_{t_2}$, we deduce that (3.22) holds for $t \in [t_2, t_3)$, and so on. ■

Note that if $m > 1$, both queues are *initially empty* and $S_1 > m\tau_1$, then $W_{t_2}^{(1)} = 0$ (because rest $m-1 \ge 1$ servers are empty) but $W_{t_2}^* = (S_1/m - \tau_1)^+ > 0$. Thus, the above established bounds are generally false for the waiting times themselves.

Now we show that FCFS (First-Come-First-Served) discipline is an optimal one in a certain sense. Consider original m-server queue Q and any m-server queue $\tilde{Q}$ with possible non-FCFS service discipline assuming initially empty systems, $\nu(0) = \tilde{\nu}(0) = 0$ and $t_1 = 0$. Also we assume that now service times S_n correspond to the order of joining service. This new coupling holds distributional properties because of the i.i.d. assumption. (As always we keep original notations for the coupled variables.) Let B_n be the nth initiation service instant and D_n be the nth service completion instant in queue Q.

Theorem 3.8. *For the coupled queues* $Q, \tilde{Q}$,

$$D_n \le \tilde{D}_n,\ n \ge 1, \tag{3.23}$$

and

$$\nu(t) \le \tilde{\nu}(t),\ t \ge 0. \tag{3.24}$$

Proof. Because $t_n = \tilde{t}_n$ and $\tilde{Q}$ is non-FCFS queue it follows from (3.11) that

$$\tilde{B}_n \ge t_n = B_n,\ n = 1, \dots, m. \tag{3.25}$$

In general, service completion epochs are defined by the following relations

$$\begin{aligned} D_n &= \Theta_n(B_1 + S_1, \dots, B_{n+m-1} + S_{n+m-1}), \\ \tilde{D}_n &= \Theta_n(\tilde{B}_1 + S_1, \dots, \tilde{B}_{n+m-1} + S_{n+m-1}),\ n \ge 1. \end{aligned} \tag{3.26}$$

Assume now that

$$\tilde{B}_i \geq B_i, \quad i = 1, \ldots, n, \tag{3.27}$$

for some $n > m$ and prove (3.27) for $i = n + 1$. By (3.11), (3.26), (3.27),

$$\begin{aligned} B_{n+1} &= \max(t_{n+1},\, D_{n+1-m}) \\ &= \max(t_{n+1},\, \Theta_{n+1-m}(B_i + S_i : i \leq n)) \\ &\leq \max(t_{n+1},\, \Theta_{n+1-m}(\tilde{B}_i + S_i : i \leq n)) \\ &= \max(t_{n+1},\, \tilde{D}_{n+1-m}) = \tilde{B}_{n+1}. \end{aligned}$$

Thus (3.27) holds for any n. It now follows from (3.26) that (3.23) also holds. Recall that $N(t) = \max(n \geq 1 : t_n \leq t)$, and that $\Lambda(t)$, $\tilde{\Lambda}(t)$ are the number of departures in $(0, t]$ in Q, $\tilde{Q}$, respectively. By (3.23), $\Lambda(t) \geq \tilde{\Lambda}(t)$, and (3.24) follows since

$$\nu(t) = N(t) - \Lambda(t) \leq N(t) - \tilde{\Lambda}(t) = \tilde{\nu}(t), \ t \geq 0.$$

■

The following statement for m-server FCFS queue Q and non-FCFS queue $\tilde{Q}$ has been established in [24, 25] but we follow [1] to present the proof. (Recall that the 1st customer arrives at instant $t_1 = 0$ at empty queues.)

Theorem 3.9. *For queues $Q, \tilde{Q}$ for each t*

$$W_t = \sum_{i=1}^{m} W_i(t) \leq_{st} \tilde{W}_t = \sum_{i=1}^{m} \tilde{W}_i(t); \tag{3.28}$$

$$\sum_{i=1}^{m} W_n^{(i)} \leq_{st} \sum_{i=1}^{m} \tilde{W}_n^{(i)}. \tag{3.29}$$

Proof. Fix $t > 0$ and introduce the number of customers

$$J(t) = \max(n \geq 1 : B_n \leq t), \ \tilde{J}(t) = \max(n \geq 1 : \tilde{B}_n \leq t),$$

which enter servers in interval $[0,\, t]$ in Q and $\tilde{Q}$, respectively. Now we will use the following modification of the coupling. For each t, we keep allocation of service times for customers who join service in $[0,\, t]$ as above (according to the order in which they enter server) and allocate remaining service times $N(t) - J(t)$, respectively $N(t) - \tilde{J}(t)$, in the FCFS order in both queues. Denote coupled queues (which keep distributions) as Q^*, $\tilde{Q}^*$, respectively. By construction (in evident notations), for each fixed t,

$$W_t^* =_{st} W_t, \ \tilde{W}_t^* =_{st} \tilde{W}_t.$$

Because $B_n \le \tilde{B}_n$, then $J(t) \ge \tilde{J}(t)$ for each t. Moreover, $B_n \le t$ for $n \le J(t)$. Thus we obtain

$$\begin{aligned} W_t^* &= \sum_{n=1}^{J(t)} (B_n + S_n - t)^+ + \sum_{n=J(t)+1}^{N(t)} S_n \\ &\le \sum_{n=1}^{\tilde{J}(t)} (\tilde{B}_n + S_n - t)^+ + \sum_{n=\tilde{J}(t)+1}^{J(t)} S_n + \sum_{n=J(t)+1}^{N(t)} S_n = \tilde{W}_t^*, \end{aligned}$$

because customers $n = \tilde{J}(t)+1, \ldots, J(t)$ are still waiting in $\tilde{Q}^*$, and thus $(\tilde{B}_n + S_n - t)^+ = S_n$ for these customers, while they have already joint service in Q^*. This coupling deals with permutations (different for Q and $\tilde{Q}$) of customers $1, \ldots, N(t)$. Thus we obtain inequality

$$W_t^* \le \tilde{W}_t^* = \sum_{i=1}^{m} \tilde{W}_i^*(t),$$

which also holds for $t = t_n^-$. ■

Unlike previous models, this result is false in the sample-path sense because we use (new) coupling for each t. By this reason we can not replace notation $\le_{st}$ by $\le$ in (3.28), (3.29).

Note 3.2. Some previous results can be extended to initially non-empty queues under appropriate initial conditions.

Note 3.3. We also mention monotonicity of a multiserver retrial queue Q with a finite buffer, batch arrivals and with exponential retrial times, established recently in [26]. Consider an initially empty queue Q with retrial rate $\gamma(n)$ depending on current number n of customers in orbit, and let $X_1(t)$ be the number of customer in queue, $X_2(t)$ be the number of customers in orbit at instant t. Denote $X(t) = (X_1(t), X_2(t))$. Let tildes denote variables in a retrial system $\tilde{Q}$ with $\tilde{\gamma}(n) \le \gamma(n)$, $n \ge 1$. In particular, it has been proved that $X(t) \le_{st} \tilde{X}(t)$ for $t \ge 0$. (We omitt the proof based on laborious path-wise comparisons.)

3.2. **Queues with dependencies.** Following [2], we now present coupling for dependent interarrival and service times $\{\tau_1, S_1, \tau_2, S_2, \ldots\}$ in a $G/G/1$ queue determined by conditional distributions

$$\begin{aligned} F_1(x) &= \mathsf{P}(\tau_1 \le x), \\ F_2(x_1, x_2) &= \mathsf{P}(S_1 \le x_2 | \tau_1 \in dx_1), \\ F_3(x_1, x_2, x_3) &= \mathsf{P}(\tau_2 \le x_3 | \tau_1 \in dx_1,\, S_1 \in dx_2), \end{aligned} \tag{3.30}$$

and so on. Note that one can use another order of interarrival and service times. Define new random variables via quantile functions on a common probability space Ω^* as follows: for any $\omega^* \in \Omega^*$,

$$\begin{aligned} \tau_1^*(\omega^*) &= \inf(x : F_1(x) \geq \omega^*); \\ S_1^*(\omega^*) &= \inf(x : F_2(\tau_1(\omega^*),\, x) \geq \omega^*), \\ \tau_2^*(\omega^*) &= \inf(x : F_3(\tau_1(\omega^*),\, S_1^*(\omega^*),\, x) \geq \omega^*), \end{aligned}$$

etc. It is obvious that $\mathsf{P}(\tau_1^* \leq x) = F_1(x)$ and $\tau_1^*(\omega^*) \leq x$ iff $\omega^* \leq F_1(x)$. Moreover,

$$S_1^*(\omega^*) \leq x \text{ if and only if } F_2(\tau_1(\omega^*),\, x) \geq \omega^*.$$

Because ω^* is uniformly distributed in $(0, 1)$,

$$\begin{aligned} \mathsf{P}(S_1^*(\omega^*) \leq x_2 | \tau_1^*(\omega^*) \in dx_1) &= \mathsf{P}(F_2(\tau_1(\omega^*),\, x_2) \geq \omega^* | \tau_1^*(\omega^*) \in dx_1) \\ &= \mathsf{P}(F_2(x_1,\, x_2) \geq \omega^*) = F_2(x_1,\, x_2), \end{aligned}$$

and so on. Thus, new random variables defined on a common probability space hold predefined (marginal) conditional distributions.

The paper [2] also extends Theorems 2.2, 3.1, 3.2, 4.1, 4.2 from [4] to a $GI/G/m$ queue with dependencies given by conditional distributions under suitable *consistency assumptions*.

4. Stochastic monotonicity in the networks

In this section, we follow a landmark paper [17]. Consider a general stochastic network with M stations, FCFS service discipline and denote for station i: the number of servers m_i, the nth external arrival instant $E_i(n)$, the nth service time $S_i(n)$, next station $V_i(n)$ to be visited by the customer that is the nth to depart station i ($V_i(n) = 0$ if the customer leaves the network), $\nu_i(t)$ the number of customers at instant t with a given initial state $\nu_i(0)$. The *input sequence*

$$\{\nu_i(0),\, E_i(n),\, S_i(n),\, V_i(n),\, i = 1, \ldots, M;\, n \geq 1\} =: \{\nu(0), \mathsf{E},\, \mathsf{S},\, \mathsf{V}\}, \quad (4.1)$$

is assumed to be given, where $\nu(0) = (\nu_1(0), \ldots, \nu_M(0))$ is initial allocation of customers. We call V *routing*. Now we define for each station i the *output sequence*

$$\{A_i(n),\, B_i(n),\, D_i(n),\, i = 1, \ldots, M;\, n \geq 1\} =: \{\mathsf{A},\, \mathsf{B},\, \mathsf{D}\}, \quad (4.2)$$

where $A_i(n)$ is the nth arrival instant, $B_i(n)$ is the nth service initiation epoch and $D_i(n)$ is the nth service completion instant. Let $1(A)$ be the indicator of an event A, and $1^{-1}(A)$ be its *reciprocal*, that is $1(A) = 1^{-1}(A) = 1$ if A is true, and $1^{-1}(A) = \infty$ if $1(A) = 0$. The following representation of

the output sequence in the terms of the input sequence is now obvious.

Theorem 4.1.

$$A_i(n) = \Theta_n\Big(E_i(m) \bigcup \Big[\bigcup_{j=1}^{M} D_{ji}(m)\Big],\, m \le n\Big), \tag{4.3}$$

where

$$D_{ji}(m) = \Theta_m\{D_j(l) \cdot 1^{-1}(V_j(l) = i),\, l = 1, 2, \ldots\}; \tag{4.4}$$

$$B_i(n) = \max(A_i(n),\, D_i(n - m_i)); \tag{4.5}$$

$$D_i(n) = \Theta_n(B_i(k) + S_i(k);\, k \le n + m_i - 1), \tag{4.6}$$

with $E_i(n) = 0$ *for* $n = -1, \ldots, -\nu_i(0)$.

Note that we use negative numbering for the customers initially presented in the network. If customer l leaving station j does not go to station i then $1^{-1}(V_j(l) = i) = \infty$ and thus $D_{ji}(m)$ is the mth internal transition instant $j \to i$.

The following statement is a direct consequence of the representation (4.3)-(4.6). In fact, the proof is based on induction in the number of transitions and the 1st step of induction must be established for *initial customers* $\nu(0)$.

Theorem 4.2. *The output sequence* $\{\mathsf{A}, \mathsf{B}, \mathsf{D}\}$ *is*
i) increasing in $\{\mathsf{E}, \mathsf{S}\}$
ii) decreasing in $\nu(0)$ *and* $\{m_i\}$.

For each station i and any interval $[0, t]$, introduce the following notations: the number of departures

$$N_i^D(t) = \sup(n : D_i(n) \le t); \tag{4.7}$$

the number of external arrivals,

$$N_i^E(t) = \sup(n : E_i(n) \le t); \tag{4.8}$$

the total number of arrivals,

$$N_i^A(t) = \sup(n : A_i(n) \le t); \tag{4.9}$$

the total number of arrivals leaving network after station i,

$$N_{i0}^D(t) = \sup(n : D_{i0}(n) \le t), \tag{4.10}$$

where

$$D_{i0}(n) = \Theta_n\Big(D_i(l) \cdot 1^{-1}(V_i(l) = 0),\, l = 1, 2, \ldots\Big) \tag{4.11}$$

is the nth instant when a customer leaves the network (after station i). Also introduce the total number of customers in the network at instant t,

$$\nu_t = \sum_{i=1}^{M} (N_i^E(t) - N_{i0}^D(t)),\ t \geq 0. \tag{4.12}$$

In what follows we will consider two stochastic networks Σ, $\tilde{\Sigma}$ (with tildes denoting variables in $\tilde{\Sigma}$). Stochastic ordering between two vectors is assumed to be component wise.

We establish a monotonicity of the processes in a closed network w. r. t. initial state $\nu(0)$. (Closed network contains M stations and a fixed number of customers $\sum_{i=1}^{M} \nu_i(0)$ circulating between stations.) Let $\nu(t) = (\nu_1(t), \ldots, \nu_M(t))$, $t \geq 0$.

Theorem 4.3. *Consider two closed networks Σ, $\tilde{\Sigma}$ with equivalent service processes and routing,*

$$\{\mathsf{S},\, \mathsf{V}\} =_{st} \{\tilde{\mathsf{S}},\, \tilde{\mathsf{V}}\},$$

which are independent of the initial states $\nu(0)$ and $\tilde{\nu}(0)$, respectively. Assume

$$\nu(0) \leq_{st} \tilde{\nu}(0). \tag{4.13}$$

Then the output sequences are ordered as

$$\{\mathsf{A},\, \mathsf{B},\, \mathsf{D}\} \geq_{st} \{\tilde{\mathsf{A}},\, \tilde{\mathsf{B}},\, \tilde{\mathsf{D}}\}, \tag{4.14}$$

and moreover

$$\{N_i^D(t),\, i = 1, \ldots, M\} \leq_{st} \{\tilde{N}_i^D(t),\, i = 1, \ldots, M\}. \tag{4.15}$$

Proof. We let

$$E_i(n) = 0,\ \ n \leq \nu_i(0),\ \ E_i(n) = \infty,\ \ n > \nu_i(0),\ \ i = 1, \ldots, M, \tag{4.16}$$

and similarly for the network $\tilde{\Sigma}$. Obviously, the networks have constant populations, which, by (4.13), satisfy inequality

$$\sum_{i=1}^{M} \nu_i(0) \leq_{st} \sum_{i=1}^{M} \tilde{\nu}_i(0).$$

On a common probability space we generate new initial data $\nu^*(0) = (\nu_1^*(0), \ldots, \nu_M^*(0))$, $\hat{\nu}(0) = (\hat{\nu}_1(0), \ldots, \hat{\nu}_M(0))$ such that

$$\nu^*(0) =_{st} \nu(0),\ \ \hat{\nu}(0) =_{st} \tilde{\nu}(0),$$

and

$$\nu^*(0) \leq \hat{\nu}(0). \tag{4.17}$$

Because $\sum_{i=1}^M \nu_i^*(0) \le \sum_{i=1}^M \hat{\nu}_i(0)$, it then follows from (4.16) that

$$\mathsf{E}^* \ge \hat{\mathsf{E}}. \tag{4.18}$$

On the same probability space we now generate sequences $(\mathsf{S}^*, \mathsf{V}^*) = (\hat{\mathsf{S}}, \hat{\mathsf{V}})$ $(=_{st} (\mathsf{S}, \mathsf{V}))$ and independent of the initial states $\nu^*(0)$, $\hat{\nu}(0)$, respectively. Then given input sequences $(\nu^*(0), \mathsf{S}^*, \mathsf{V}^*)$, $(\hat{\nu}(0), \hat{\mathsf{S}}, \hat{\mathsf{V}})$, using representation (4.3)-(4.6) one constructs the output sequences $\{\mathsf{A}^*, \mathsf{B}^*, \mathsf{D}^*\}$, $\{\hat{\mathsf{A}}, \hat{\mathsf{B}}, \hat{\mathsf{D}}\}$ in such a way that

$$\{\mathsf{A}^*, \mathsf{B}^*, \mathsf{D}^*\} =_{st} \{\mathsf{A}, \mathsf{B}, \mathsf{D}\}, \ \ \{\hat{\mathsf{A}}, \hat{\mathsf{B}}, \hat{\mathsf{D}}\} =_{st} \{\tilde{\mathsf{A}}, \tilde{\mathsf{B}}, \tilde{\mathsf{D}}\}. \tag{4.19}$$

Then it follows from (4.17), (4.18) and Theorem 4.2 (i), that

$$\{\mathsf{A}^*, \mathsf{B}^*, \mathsf{D}^*\} \ge \{\hat{\mathsf{A}}, \hat{\mathsf{B}}, \hat{\mathsf{D}}\}. \tag{4.20}$$

Thus (4.14) is proved. To prove (4.15), we apply definitions (4.7)-(4.11). Indeed, $A_i^*(n) \ge \hat{A}_i(n)$ implies $B_i^*(n) \ge \hat{B}_i(n)$ and the latter implies $D_i^*(n) \ge \hat{D}_i(n)$ and $D_{i0}^*(n) \ge \hat{D}_{i0}(n)$ because we use the same routing (coupling) and $1^{-1}(V_j(l) = i) = 1$ if and only if $1^{-1}(\tilde{V}_j(l) = i) = 1$ for all i, j, l. Thus,

$$(N_i^{*D}(t),\, i = 1, \dots, M) \le_{st} (\hat{N}_i^D(t),\, i = 1, \dots, M), \tag{4.21}$$

and (4.15) follows. ■

Now we establish the similar monotonicity property of the network processes (both for open and closed networks) w.r.t. service times.

Theorem 4.4 *Consider two open (or closed) networks* Σ, $\tilde{\Sigma}$ *with service times sequences*

$$\mathsf{S} \ge_{st} \tilde{\mathsf{S}},$$

which are independent of the routing, external input and initial states, and assume that

$$\{\nu(0), \mathsf{E}, \mathsf{V}\} =_{st} \{\tilde{\nu}(0), \tilde{\mathsf{E}}, \tilde{\mathsf{V}}\}.$$

Then

$$\{\mathsf{A}, \mathsf{B}, \mathsf{D}\} \ge_{st} \{\tilde{\mathsf{A}}, \tilde{\mathsf{B}}, \tilde{\mathsf{D}}\}; \tag{4.22}$$

$$(N_i^D(t),\, i = 1, \dots, M) \le_{st} (\tilde{N}_i^D(t),\, i = 1, \dots, M); \tag{4.23}$$

and for an open network also

$$(N_{i0}^D(t),\, i = 1, \dots, M) \le_{st} (\tilde{N}_{i0}^D(t),\, i = 1, \dots, M). \tag{4.24}$$

Proof. Formulas (4.22), (4.23) are deduced as (4.20), (4.21) since, by Theorem 4.2, the output sequence is increasing in service times. Since we use the same routing, the sequence $\{D_{i0}(n)\}$ is monotone (see (4.10), (4.11)), and (4.24) follows. ■

Previous results give the following statement.

Corollary. *Consider two open networks* Σ, $\tilde{\Sigma}$ *with input and service times sequences*

$$\{\mathsf{E},\, \mathsf{S}\} \geq_{st} \{\tilde{\mathsf{E}},\, \tilde{\mathsf{S}}\},$$

independent of the initial states

$$\{\nu(0),\, \mathsf{V}\} =_{st} \{\tilde{\nu}(0),\, \tilde{\mathsf{V}}\}.$$

Then (4.22), (4.24) hold.

Theorem 4.5. *Assume that in two networks* Σ, $\tilde{\Sigma}$ *(open or closed) the number of servers at each station* i *are connected as*

$$m_i \leq \tilde{m}_i,\ i = 1, \ldots, M.$$

Then the statement of Theorem 4.4 holds.

Proof. It follows from (4.5) that $B_i(n)$ decreases in m_i. By (4.6), decreasing $B_i(n)$ implies increasing $D_i(n)$. Finally, arrival sequence $A_i(n)$ decreases in $D_{ji}(n)$. ■

5. Regenerative stability analysis

In this Section, we present a few examples of stability analysis of regenerative systems using several new and also above obtained monotonicity results.

First, we consider a 2-station tandem network $GI/G/1 \to \cdot \to \cdot/G/1$ where the input to station 1 is rejected if the queue size at station 2 exceeds a predetermined threshold $N > 0$. We call it *feedback admission control*. The (external) input to station 1 is a stationary renewal process with instants t_n^*, $n \geq 1$, and rate λ. Let $S_n^{(i)}$, $n \geq 1$, be the i.i.d. service times at station i with rate μ_i and let $\nu_i(t)$ be the queue size at station i at instant t^-, $i = 1, 2$; $t \geq 0$. Thus, the input to station 1 is rejected as long as $\nu_2(t) \geq N$. Such a model (under exponential assumptions) has been introduced in [12, 13]. Moreover, stability analysis of this network is also discussed in [15, 16]. For this network, we establish a few new monotonicity properties and, on this basis, develop stability analysis when $\mu_1 > \mu_2$.

We exploit a regenerative structure of the (right-continuous) basic queue-size process, $\nu(t) = (\nu_1(t),\, \nu_2(t))$, $t \geq 0$, which is classically regenerative with regeneration instants

$$\beta_0 = 0,\ \beta_{n+1} = \inf\Big(t_k^* > \beta_n :\ \nu(t_k^*) = (0,\, 0)\Big),\ n \geq 0, \tag{5.1}$$

and the i.i.d. regeneration periods $\alpha_n = \beta_{n+1} - \beta_n$, $n \geq 1$, independent of β_1. Of course, (5.1) are classical regenerations of the one-dimensional processes $\nu_i(t)$, $t \geq 0$, $i = 1, 2$. We use the following characterization of the forward regeneration time $\beta(t) = \min\{\beta_k - t : \beta_k - t > 0\}$ at instant $t \geq 0$, see [3]. We assume $\beta_1 < \infty$ with probability 1 (w.p.1). Then the mean regeneration period $\mathsf{E}\alpha = \infty$ if and only if

$$\beta(t) \to \infty \ \text{ in probability as } \ t \to \infty. \tag{5.2}$$

If $\mathsf{E}\alpha < \infty$ then the process $\{\beta(t), t \geq 0\}$ is tight and renewal process of regenerations (5.1) and the basic queueing process are called *positive recurrent*. To establish positive recurrence (the key element of stability), we show that (5.2) does not hold.

Denote original 2nd station Q (with the queue-size $\nu_2(t)$), and also consider a modified system $\tilde{Q}$ (with queue-size $\tilde{\nu}_2(t)$) which is fed by the 1st saturated station of original network Q. (Again we apply a coupling using the sample-path equivalence between renewal output process from the 1st station and the renewal input to $\tilde{Q}$.)

More exactly, interarrival times in $\tilde{Q}$ are service times $\{S_n^{(1)}\}$. In what follows tildes denote the variables describing system $\tilde{Q}$. Let, for station Q, t_n be the nth input instant, W_n be the waiting time of customer n and let T_n be the nth output instant, $n \geq 1$. We assume the same initial states in both queues Q, $\tilde{Q}$. More precisely, if the 1st station (in the tandem) is busy at instant $t = 0$ with unfinished service time $S_0^{(1)}$, then $S_0^{(1)}$ is also the 1st input interval in $\tilde{Q}$. Otherwise, the 1st input interval in $\tilde{Q}$ equals $\tilde{t}_1 = S_1^{(1)}$, while at the 1st station, an empty period $\mu_0 \geq 0$ precedes first service and thus, $t_1 = S_1^{(1)} + \mu_0 \geq \tilde{t}_1$. Hence, $\tilde{t}_1 = S_0^{(1)} I_{\nu_1(0)>0} + S_1^{(1)} I_{\nu_1(0)=0}$. It follows from construction that the difference $\mu_{n-1} := t_n - \tilde{t}_n \geq 0$ is the amount of an empty time at the 1st queue after the nth departure, $n \geq 1$. Denote $\mu(n) = \sum_{k=0}^{n-1} \mu_k$ and introduce input intervals $\tau_n = t_{n+1} - t_n$, $\tilde{\tau}_n = \tilde{t}_{n+1} - \tilde{t}_n$, $n \geq 1$. It is also assumed that initial unfinished workload in both queues Q, $\tilde{Q}$ has the same value x_0. By construction, $\tilde{t}_1 + \mu_0 = t_1$ and for $n \geq 1$,

$$\tilde{t}_n = \sum_{k=1}^{n} S_k^{(1)}, \ \text{if } \nu_1(0) = 0; \quad \tilde{t}_n = \sum_{k=0}^{n-1} S_k^{(1)}, \ \text{if } \nu_1(0) > 0. \tag{5.3}$$

Moreover,

$$t_{n+1} = t_n + S_{n+1}^{(1)} + \mu_n, \ t_n = \tilde{t}_n + \mu(n), \ n \geq 1. \tag{5.4}$$

Theorem 5.1.

$$\tilde{T}_n \leq T_n \leq \tilde{T}_n + \mu(n), \ \ W_n \leq \tilde{W}_n \leq W_n + \mu(n), \ n \geq 1. \tag{5.5}$$

Proof. In what follows we use coupling taking $S_n^{(2)}$ as the service time of customer n in both queues $Q, \tilde{Q}$. Also recall that a coupling is used to generate input to stations Q and $\tilde{Q}$ taking the same service times $S_n^{(1)}$ according to (5.4). Obviously,

$$W_1 = (x_0 - t_1)^+ \le (x_0 - \tilde{t}_1)^+ = \tilde{W}_1 \le \mu_0 + (x_0 - \tilde{t}_1 - \mu_0)^+ = W_1 + \mu(1). \quad (5.6)$$

It then follows that

$$\tilde{T}_1 = \tilde{t}_1 + \tilde{W}_1 + S_1^{(2)} = \tilde{t}_1 + \mu_0 + \tilde{W}_1 - \mu_0 + S_1^{(2)} \le t_1 + W_1 + S_1^{(2)} = T_1, \quad (5.7)$$

and by (5.6),

$$T_1 = \tilde{t}_1 + \mu_0 + W_1 + S_1^{(2)} \le \tilde{T}_1 + \mu(1). \quad (5.8)$$

Based on (5.6)–(5.8), we prove (5.5) for $n+1$ by induction, assuming it holds for $k = 1, \dots, n$. Because $\tilde{\tau}_n = S_n^{(2)}$ and $\tau_n = S_n^{(2)} + \mu_{n-1}$, $n \ge 1$, then by induction assumption,

$$\begin{aligned} \tilde{T}_{n+1} &= \tilde{t}_{n+1} + (\tilde{W}_n + S_n^{(2)} - S_n^{(1)})^+ + S_{n+1}^{(2)} \\ &\le t_{n+1} + (\tilde{W}_n - \mu(n) + S_n^{(2)} - S_n^{(1)} - \mu_n)^+ + S_{n+1}^{(2)}, \\ &\le t_{n+1} + (W_n + S_n^{(2)} - S_n^{(1)} - \mu_n)^+ + S_{n+1}^{(2)} = T_{n+1}. \end{aligned}$$

On the other hand (again by induction assumption),

$$\begin{aligned} T_{n+1} &= t_{n+1} + (W_n + S_n^{(2)} - S_n^{(1)} - \mu_n)^+ + S_{n+1}^{(2)} \\ &\le \tilde{t}_{n+1} + \mu(n+1) + (\tilde{W}_n + S_n^{(2)} - S_n^{(1)})^+ + S_{n+1}^{(2)} \\ &= \tilde{T}_{n+1} + \mu(n+1), \end{aligned}$$

and the 2nd group of inequalities in (5.5) follows. To estimate workload, we have

$$\begin{aligned} W_{n+1} &= (W_n + S_n^{(2)} - S_n^{(1)} - \mu_n)^+ \le (\tilde{W}_n + S_n^{(2)} - S_n^{(1)})^+ = \tilde{W}_{n+1} \\ &\le (W_n + \mu(n) + S_n^{(2)} - S_n^{(1)} - \mu_n)^+ \\ &\le (W_n + S_n^{(2)} - S_n^{(1)} - \mu_n)^+ + \mu(n+1) \\ &= W_{n+1} + \mu(n+1). \end{aligned}$$

where we use inequality $\mu(n) \le \mu(n+1)$. Thus, (5.5) is proved for all n. ■

Note 5.1. Because $t_n \ge \tilde{t}_n$, then inequality $T_n \ge \tilde{T}_n$ also follows from [7, 9, 17].

Note 5.2. It follows from the Kiefer -Wolfowitz representation (3.2) that monotonicity (5.5) also holds for the waiting times $\{W_n^{(1)}\}$ in a $GI/G/m$ system, that is

$$W_n^{(1)} \le \tilde{W}_n^{(1)} \le W_n^{(1)} + \mu(n),\ n \ge 1. \quad (5.9)$$

The 2nd inequality follows as in the proof above, and the 1st inequality (5.9) holds since $\tau_n - \tau'_n \geq \mu_{n-1} + \mu_n \geq 0$ and we can apply Theorem 3.1 and the monotonicity of mapping (3.2).

Now we establish a reduction of the total time when the queue size (in Q) exceeds any fixed threshold k, in comparison with the time in queue $\tilde{Q}$, within any interval $[0,\, t]$. Fix some n for a moment and denote

$$a_1 = T_n - t_{n+k-1}, \;\; \tilde{a}_1 = \tilde{T}_n - \tilde{t}_{n+k-1}, \;\; a_2 = a_1 - S_n^{(2)}, \;\; \tilde{a}_2 = \tilde{a}_1 - S_n^{(2)}.$$

Note that the nth beginning service time (in Q) is defined as $B_n = \max(t_n, T_{n-1}) = t_n + W_n$. Define

$$\tilde{l}_n = \tilde{a}_1^+ - \tilde{a}_2^+, \; l_n = a_1^+ - a_2^+, \tag{5.10}$$

and note that $\tilde{l}_n$ and l_n are the parts of the nth service intervals

$$[\tilde{B}_n,\, \tilde{B}_n + S_n^{(2)}) \;\text{ and }\; [B_n,\, B_n + S_n^{(2)}),$$

respectively, such that $\tilde{\nu}_2(t) \geq k$ and $\nu_2(t) \geq k$, see [1].

Theorem 5.2.

$$l_n \leq l'_n, \; n \geq 1. \tag{5.11}$$

Proof. Note that $a_1 \geq a_2$, $\tilde{a}_1 \geq \tilde{a}_2$.

i) Let $\tilde{a}_1 > 0$, $\tilde{a}_2 > 0$. Then $l_n = S_n^{(2)}$. If moreover, $a_1 > 0$, $a_2 > 0$, then $\tilde{l}_n = l_n$. Now we assume $a_1 > 0$, $a_2 \leq 0$. Then $l_n = T_n - t_{n+k-1} \leq S_n^{(2)} = \tilde{l}_n$.

ii) Let $\tilde{a}_1 > 0$, $\tilde{a}_2 \leq 0$. Then $0 < \tilde{l}_n = \tilde{T}_n - \tilde{t}_{n+k-1} \leq S_n^{(2)}$. Denote $\Delta_n(k) = \mu(n+k-1) - \mu(n)$, $k \geq 1$. Because $\Delta_n(k) \geq 0$ then it follows from (5.5) that

$$\begin{aligned} a_2 &= T_n - S_n^{(2)} - \tilde{t}_{n+k-1} - \mu(n+k-1) \\ &= T_n - \mu(n) - S_n^{(2)} - \tilde{t}_{n+k-1} - \Delta_n(k) \\ &\leq \tilde{T}_n - S_n^{(2)} - \tilde{t}_{n+k-1} = \tilde{a}_2 \leq 0. \end{aligned}$$

Thus,

$$l_n = a_1^+ \leq (T_n - \mu(n) - \tilde{t}_{n+k-1})^+ \leq (\tilde{T}_n - \tilde{t}_{n+k-1})^+ = \tilde{a}_1^+ = \tilde{l}_n.$$

iii) Eventually, let $\tilde{a}_1 \leq 0$. Then $\tilde{a}_2 \leq 0$ and thus, $\tilde{l}_n = 0$. Since $\tilde{T}_n \leq \tilde{t}_{n+k-1}$, we obtain

$$a_1 = \tilde{T}_n - t_{n+k-1} - \mu(n+k-1) \leq \tilde{T}_n - \tilde{t}_{n+k-1} - \Delta_n(k) \leq 0.$$

Because also $a_2 \leq 0$ then $l_n = 0$, and (5.11) follows. ■

Denote

$$\tilde{\mu}_k(t) = \int_0^t I_{\tilde{\nu}_2(s) \geq k} ds, \; \mu_k(t) = \int_0^t I_{\nu_2(s) \geq k} ds,$$

the time when queues in $\tilde{Q}$, Q, respectively exceeds a (fixed) threshold k (within interval $[0,\, t]$). The following monotonicity property holds.

Corollary 5.1.

$$\tilde{\mu}_k(t) \geq \mu_k(t),\ t \geq 0. \tag{5.12}$$

Proof. Because $t_n \geq \tilde{t}_n$, $T_n \geq \tilde{T}_n$, it then follows that the service beginning times are connected as $B_n \geq \tilde{B}$ and hence,

$$N(t) = \#\{n : B_n \leq t\} \leq \#\{n : \tilde{B}_n \leq t\} = \tilde{N}(t),\ t \geq 0.$$

Now we obtain from (5.11) that (5.12) holds for each t and any threshold k:

$$\tilde{\mu}_k(t) = \sum_{k=1}^{\tilde{N}(t)} \tilde{l}_k \geq \sum_{k=1}^{N(t)} l_k = \mu_k(t),\ t \geq 0.$$

■

Now we apply previous results to stability analysis of original tandem network with feedback admission control. (More details on stability analysis using characterization (5.2) can be found in [10].) Let τ be a generic interarrival time.

Theorem 5.3. *Assume that $\mu_1 > \mu_2$ and condition*

$$\mathsf{P}(\tau > S_1^{(1)} + S_1^{(2)}) > 0 \tag{5.13}$$

holds. Then the processes $\{\nu(t),\, t \geq 0\}$, and $\{\nu_i(t),\, t \geq 0\}$, $i = 1, 2$, with regenerations (5.1), are positive recurrent for any $\lambda < \infty$.

Proof. Assume that

$$\nu_1(t) \to \infty \quad \text{in probability as } t \to \infty, \tag{5.14}$$

and denote the empty time of the 1st station in interval $[0, t]$ as

$$\mu_0^{(1)}(t) = \int_0^t I_{\nu_1(s)=0} ds.$$

It then follows from (5.14) that $\mathsf{P}(\nu_1(t) = 0) \to 0$, and the mean empty time is

$$\mathsf{E}\mu_0^{(1)}(t) = o(t) \text{ as } t \to \infty. \tag{5.15}$$

Let $a_1(t)\,(b_1(t))$ be the number of arrivals to (departures from) station 1 in the original network, and let $l(t)$ be the number of the rejected arrivals in interval $(0, t]$. Obviously,

$$b_1(t) + \nu_1(t) = \nu_1(0) + a_1(t) - l(t). \tag{5.16}$$

Based on the monotonicity results proved above, we now obtain bounds for the queue-size processes in Q and $\tilde{Q}$. Let $a_2(t)\,(b_2(t))$ be the number of arrivals to (departures from) station Q in $(0,\, t]$. (As above, tildes denote

variables for station $\tilde{Q}$.) By construction, the number $\tilde{a}_2(t-\mu_0^{(1)}(t))$ of arrivals to $\tilde{Q}$ in interval $(0,\, t-\mu_0^{(1)}(t)]$ is the same as the number of arrivals $a_2(t)$ to Q in interval $(0,\, t]$, that is $\tilde{a}_2(t-\mu_0^{(1)}(t)) = a_2(t)$. By the inequality (5.5), $T_n \le \tilde{T}_n + \mu(n)$, and thus the number of departures from $\tilde{Q}$ in interval $(0,\, t]$ *among* $\tilde{a}_2(t-\mu_0^{(1)}(t))$ *arrivals* is not less than the number of departures from Q within the same interval $(0,\, t]$. Assume that all arrivals to $\tilde{Q}$ in interval $(t-\mu_0^{(1)}(t), t]$ are rejected, and denote such a queue (at instant t) as $\eta_2(t)$. Then it follows that $\nu_2(t) \ge \eta_2(t)$ and moreover, $\tilde{\nu}_2(t) \le \eta_2(t) + \tilde{a}_0(t)$, where $\tilde{a}_0(t)$ is the number of new arrivals to station $\tilde{Q}$ in interval $[t-\mu_0^{(1)}(t), t]$ assuming that $t-\mu_0^{(1)}(t)$ is an arrival instant. (That is the process $\tilde{a}_0(\cdot)$ is the zero-delayed version of the input $\tilde{a}_2(\cdot)$.) It gives the following inequalities

$$\nu_2(t) \ge \eta_2(t) \ge \tilde{\nu}_2(t) - \tilde{a}_0(t),\ t \ge 0. \tag{5.17}$$

It is easy to show under assumption (5.14) that

$$\mathsf{E}\tilde{a}_0(t) = o(t) \ \text{ as } \ t \to \infty. \tag{5.18}$$

To prove it, we use an upper linear bound of the renewal function. Moreover, we apply stationarity of the renewal input to show that the number of arrivals in interval $[t-\mu_0^{(1)}(t), t]$ depends on its length $\mu_0^{(1)}(t)$ only. It follows from (5.18) that the family $\{\tilde{a}_0(t)/t,\, t>0\}$ is uniformly integrable, and thus $\tilde{a}_0(t)/t \to 0$ in probability. Because $\tilde{\nu}_2(t)/t \to \mu_1 - \mu_2 =: d > 0$ w.p.1, hence in probability, it then follows from (5.17) that $\nu_2(t)/t \to d$ in probability. In particular, $\nu_2(t) \to \infty$ in probability. Now we fix any $\delta > 0$. Then there exists $t' \ge t_0$ such that $\mathsf{P}(\nu_2(t) \ge N) \ge 1-\delta,\ t \ge t'$, and the expected fraction of the time when queue size $\nu_2(t)$ exceeds threshold N,

$$\frac{1}{t}\mathsf{E}\mu_N(t) = \frac{1}{t}\int_0^t P(\nu_2(s) \ge N)ds \to 1-\delta.$$

Since δ is arbitrary, $\mathsf{E}\mu_N(t)/t \to 1$. It now follows that the number $l(t)$ of rejected arrivals at station 1 is such that

$$\lim_{t\to\infty}\frac{1}{t}\mathsf{E}l(t) = \lim_{t\to\infty}\frac{1}{t}\mathsf{E}a_1(t).$$

Because $b_1(t) \le a_1(t) - l(t) + o(t)$ as $t \to \infty$, it follows from (5.16) that $\lim_{t\to\infty}\mathsf{E}b_1(t)/t = 0$ while (5.14) implies

$$\lim_{t\to\infty}\frac{1}{t}\mathsf{E}b_1(t) = \mu_1 > 0. \tag{5.19}$$

(More details see in [8, 9].) This contradiction shows that (5.14) does not hold. Hence,

$$\inf_i \mathsf{P}(\nu_1(z_i) = 0) \ge \varepsilon, \tag{5.20}$$

for some $\varepsilon > 0$ and a non-random sequence of instants $z_i \to \infty$. We note that if (for a fixed i) $\nu_2(z_i) \le N$, then

$$\mathsf{P}(\nu_1(z_i) = 0,\, \nu_2(z_i) \le N) \ge \varepsilon. \tag{5.21}$$

Otherwise, there exists an interval $(z_i,\, u_i]$ such that $\nu_1(t) = 0$, $t \in [z_i,\, u_i]$ and $\nu_2(u_i) \le N$ because station 1 is blocked as long as queue-size in Q exceeds threshold N. It is easy to show that on the event $\{\nu_1(z_i) = 0,\, \nu_2(z_i) \le N\}$, the lengths of intervals $u_i - z_i$ are uniformly (in i) bounded by a finite constant with a positive probability. Note that although the length $u_i - z_i$ may depend on $\nu_2(z_i)$ the following uniform lower bound for the probability holds:

$$\inf_i \mathsf{P}\Big(\nu_1(s) = 0,\, s \in (z_i,\, u_i],\, \nu_2(u_i) \le N\Big) \ge \varepsilon. \tag{5.22}$$

Thus, we conclude that there exists a non-random sequence $u_i \to \infty$ such that

$$\inf_i \mathsf{P}\Big(\nu_1(u_i) = 0,\, \nu_2(u_i) \le N\Big) \ge \varepsilon. \tag{5.23}$$

Now we apply the tightness of the unfinished interarrival time process $\tau(t) = \inf(t_n - t : t_n - t > 0)$, $t \ge 0$, and the unfinished service time process at station Q (see [7]). It then follows from (5.23) that

$$\inf_i \mathsf{P}\Big(\nu_1(u_i) = 0,\, W_2(u_i) \le R,\, \tau(u_i) \le D\Big) \ge \varepsilon/2 \tag{5.24}$$

for suitable finite constants R, D. (Note that if follows from $\nu_2(u_i) \le N$ that workload $W_2(t)$ at station 2 at instant t is bounded, $W_2(t) \le D$.) Note that (5.13) and finiteness of $\mathsf{E}\tau = 1/\lambda$ imply

$$\mathsf{P}(c \ge \tau > S_1^{(1)} + S_1^{(2)} + \varepsilon_0) \ge \delta_0 \tag{5.25}$$

for some constants $c < \infty$, $\varepsilon_0 > 0$, $\delta_0 > 0$. Denote integer part $B = \lceil R/\varepsilon_0 \rceil (\ge R/\varepsilon_0)$, $n_i = \inf(n : t_n \ge u_i)$ and, on the event $\{\nu_1(u_i) = 0,\, W_2(u_i) \le R,\, \tau(u_i) \le D\}$, realize the event

$$\bigcap_{k=1}^{B} \{\tau_{n_i+k} > S_{n_i+k}^{(1)} + S_{n_i+k}^{(2)} + \varepsilon_0\}.$$

Then a customer arrives within interval $[u_i,\, u_i + D + cB]$ which meets an empty network. Moreover, this occurs with a probability which is uniformly lower bounded by positive constant $\varepsilon\delta_0^B/2$. Thus (5.2) does not hold, and positive recurrence of regenerations (5.1) follows. ■

Keeping the same notations, we consider previous network *with no admission control* under assumption $\lambda > \mu_1$ (1st station is overloaded). Then w.p.1,

$$\nu_1(t) = \nu_1(0) + a_1(t) - b_1(t) \ge a_1(t) - \tilde{a}_2(t) + o(t) \;\; (t \to \infty),$$

where $\tilde{a}_2(t)$ is the number of renewals in $[0,\, t]$ in the process generated by service times $S_n^{(1)}$, $n \geq 1$. Hence, $\tilde{a}_2(t) \geq b_1(t)$. By assumptions,

$$\frac{a_1(t) - \tilde{a}_2(t)}{t} \to \lambda - \mu_1 > 0$$

w.p.1 as $t \to \infty$. Hence,

$$\liminf_{t\to\infty} \frac{\nu_1(t)}{t} > 0,$$

and the 1st queue in the network is *strongly unstable*. In particular, there exists a (finite) random instant T such that $\nu_1(t) > 0$ w.p.1 for all $t \geq T$. In other words, the input to the 2nd station is coupled within a finite interval with a renewal input generated by the service times $\{S_n^{(1)}\}$. Thus, we may treat the input to station 2 as a delayed renewal input (with the delay T) and with rate μ_1. Since $\mu_1 > \mu_2$ then the queue-size process $\nu_2(t)$, $t \geq 0$, is also strongly unstable.

Note 5.3. Assume that $\mu_2 > \mu_1$. Then it is easy to obtain (using coupling mentioned above) that the queue-size process $\nu_2(t)$, $t \geq 0$, is positive recurrent regenerative. If moreover $\lambda > \mu_2$, then the latter result seems surprising. Really, in this case the well-known balance equations for the (potential) input rates give the traffic intensities $\rho_1 = \lambda/\mu_1 > 1$, $\rho_2 = \lambda/\mu_2 > 1$. It may indicate an instability of both stations. Because the input to station 2 is (delayed) renewal with rate μ_1 (see above) then in fact the actual intensity for the 2nd station is $\mu_1/\mu_2 < 1$, and thus the 2nd station is not overloaded. (For more details see [8].)

ACKNOWLEDGEMENTS

I thank my colleagues Gerardo Sanz and Javier Lopez (University of Zaragoza) who have stimulated my interest in applications of the coupling theory to monotonicity of queues, and for their informative discussions. I also thank the staff of CRM for nice hospitality and remarkable conditions during my visit in September 2007. Special thanks to Rosario Delgado (UAB) for a careful reading of the draft of the paper and the numerous comments which have improved the presentation of the work.

References

[1] S. Asmussen (2003) *Applied Probability and Queues, 2nd ed.*, Springer, NY.
[2] G.L. O'Brien (1975) Inequalities for queues with dependent interarrival and service times, J.Appl.Prob. 12, 653-656,
[3] W. Feller (1971) *An Introduction to Probability Theory and its Applications*, Vol.2 (2nd edn), Wiley, NY.
[4] D.R. Jacobs and S. Schach (1972) Stochastic order relationships between $GI/G/k$ queues, Ann. Math. Stat. 43, 1623-1633.

[5] V. Kalashnikov (1994) *Topics on Regenerative Processes*, CRC Press, Roca Baton.
[6] T. Lindvall (1992) *Lectures on the Coupling Method*, Wiley.
[7] E. Morozov (1997) The tightness in the ergodic analysis of regenerative queueing processes, Queueing Systems, 27, 179-203.
[8] E. Morozov, Instability of nonhomogeneous queueing networks, Journal of Mathematical Sciences 112 (2002) 4155-4167.
[9] E. Morozov (2002). Stability of Jackson-type network output, Queueing Systems, 40, 383-406.
[10] E. Morozov (2004) Weak regeneration in modeling of queueing processes, Queueing Systems, 46, 295-315.
[11] A. Müller and D. Stoyan (2000) *Comparisons methods for stochastic models*, J. Wiley and Sons
[12] Lasse Leskelä (2004) Stabilization of an overloaded queueing network using measurement-based admission control, Helsinki University of Technology Institute of Mathematics, Research Report A470.
[13] L. Leskelä and J. Resing (2005) A Tandem Queueing Network with Feedback Admission Control, Report No 09, 2004/2005 fall, ISSN 1103-467X, ISRN IML -R –09-4/05–SE+fall.
[14] E.Morozov (1997) The stability of non-homogeneous queueing system with regenerative input, J.Math.Sci., 89, 407-421.
[15] E. Morozov (2005) Stability of a tandem network with feedback admission control, Proceedings of the 5th St.-Petersburg Workshop on Simulation, June 2005, 509-514.
[16] E. Morozov (2005) Stability of a tandem network with indirect feedback admission control, Transactions of the XXV International Seminar on Stability Problems for Stochastic Models, Majori/Salerno, University of Salerno, September 20-24, 202-204.
[17] J. Shantikumar and David Yao (1989) Stochsatic monotonicity in general queueing networks, J.Appl.Prob., 26, 413-417.
[18] K. Sigman (1990) One-dependent regenerative processes and queues in continuous time, Math. Oper. Res., 15, 175-189.
[19] C. Stone (1966) On absolutely continuous distributions and renewal theory. Ann.Math. Statist., 37, 271-275.
[20] L.Takacs (1962) *Introduction to the theory of queues*, Oxford University Press.
[21] L.Takacs (1963) The limiting distribution of the virtual waiting time and the queue size for a single-server queue with recurrent inpurt and general service time, *Sankhya*, Ser A. 25, 91-100.
[22] H. Thorrison (2000) *Coupling, Stationarity, and Regeneration*, Springer, New Work.
[23] R.W. Wolff (1977) An upper bound for multi-channel queues. *J. Appl. Probab.* 14, 884-888.
[24] R.W. Wolff (1988) Upper bounds on work in systems for multi-channel queue. *J. Appl. Probab.* 24, 547-551.
[25] R.W. Wolff (1989) *Stochastic Modeling and the Theory of Queues*, Prentice-Hall.
[26] Yang Woo Shin (2006) Monotonicty properties in various retrial queues and their applications, Queueing Systems, 53, 147-157.

Institute for Applied Mathematical Research, Karelian Research Centre RAS, and Petrozavodsk University